



# Développement d'un système d'aide au diagnostic moléculaire de maladies génétiques rares par un modèle de langage à grande échelle (LLM)\*

Recherche à l'Unité de Bio-informatique médicale appliquée au diagnostic  
NHC Strasbourg, France.

Encadrants : Jean Muller, Samuel Nicaise, Jean-Baptiste Lamouche

\*Article en cours. El alaoua et al.  
Manuscrit in preparation



**7000**

Maladies Génétiques  
Rares connues.

**10%**

Des décès entre  
1 et 5 ans.

**215**

Laboratoire de  
génétique Français.

Les maladies génétiques rares touchent **3 millions de Français.**



# Processus diagnostique

Parcours d'un patient atteint d'une  
maladie génétique rare



# Processus diagnostique

Parcours d'un patient atteint d'une  
maladie génétique rare



## Diagnostic clinique

Liste des critères cliniques  
Hypothèse diagnostique

## Diagnostic moléculaire

Séquençage de l'ADN  
Confirmation de l'hypothèse



# Objectifs de la Recherche

Développer un système d'aide au diagnostic

## I

Evaluation de l'utilisation de  
l'intelligence artificielle pour le  
diagnostic génétique

## II

Appliquer l'intelligence artificielle à  
une maladie rare



# Syndrome de Bardet-Biedl.

Ciliopathie rare (1/150 000)

- Maladie autosomique récessive.
- Hétérogène génétique avec **26 gènes BBS**
- Hétérogène clinique avec des **signes cliniques majeurs & mineurs**:

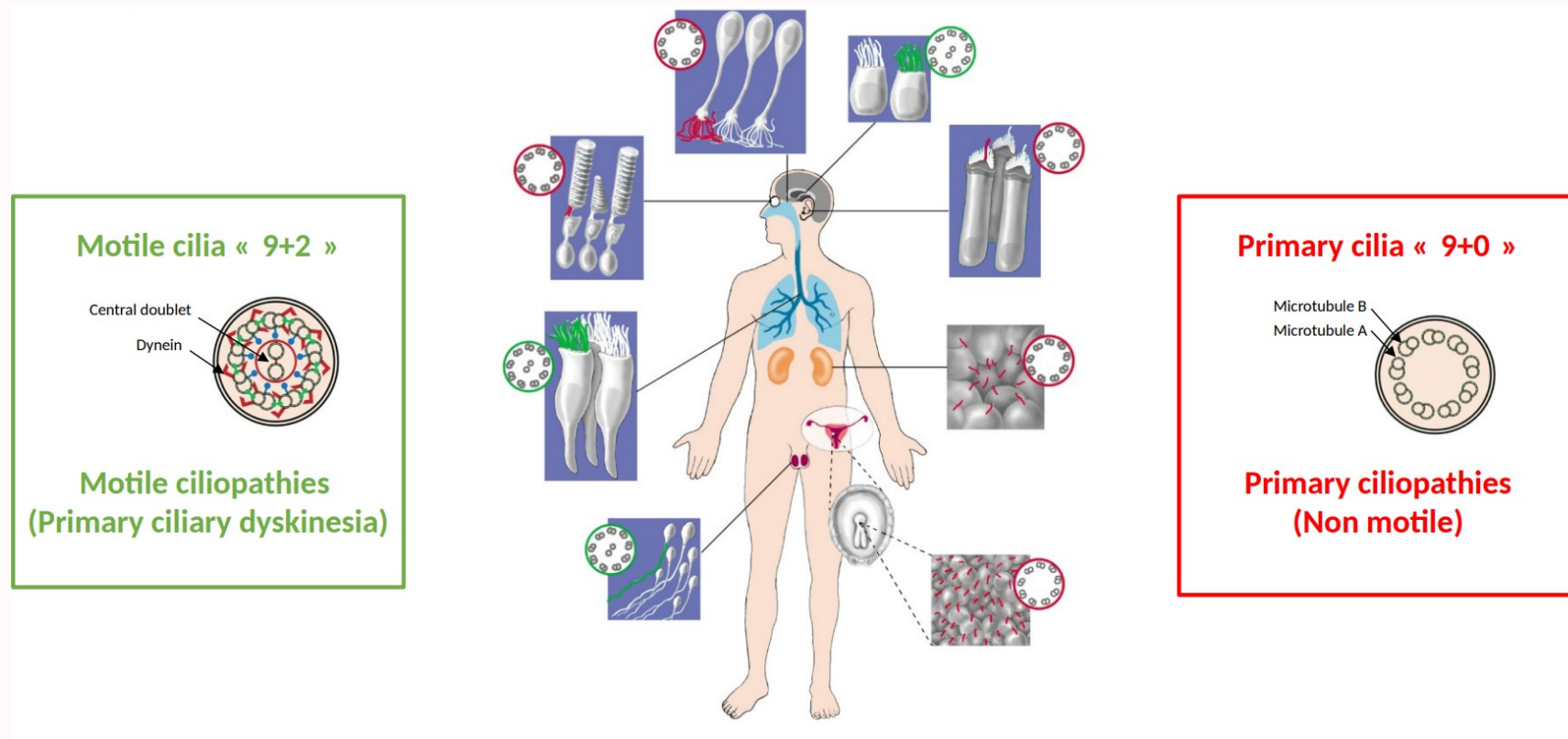
*Dystrophie rétinienne, polydactylie, atteinte rénale, obésité (critères de beales).*



# Syndrome de Bardet-Biedl.

Ciliopathie rare (1/150 000)

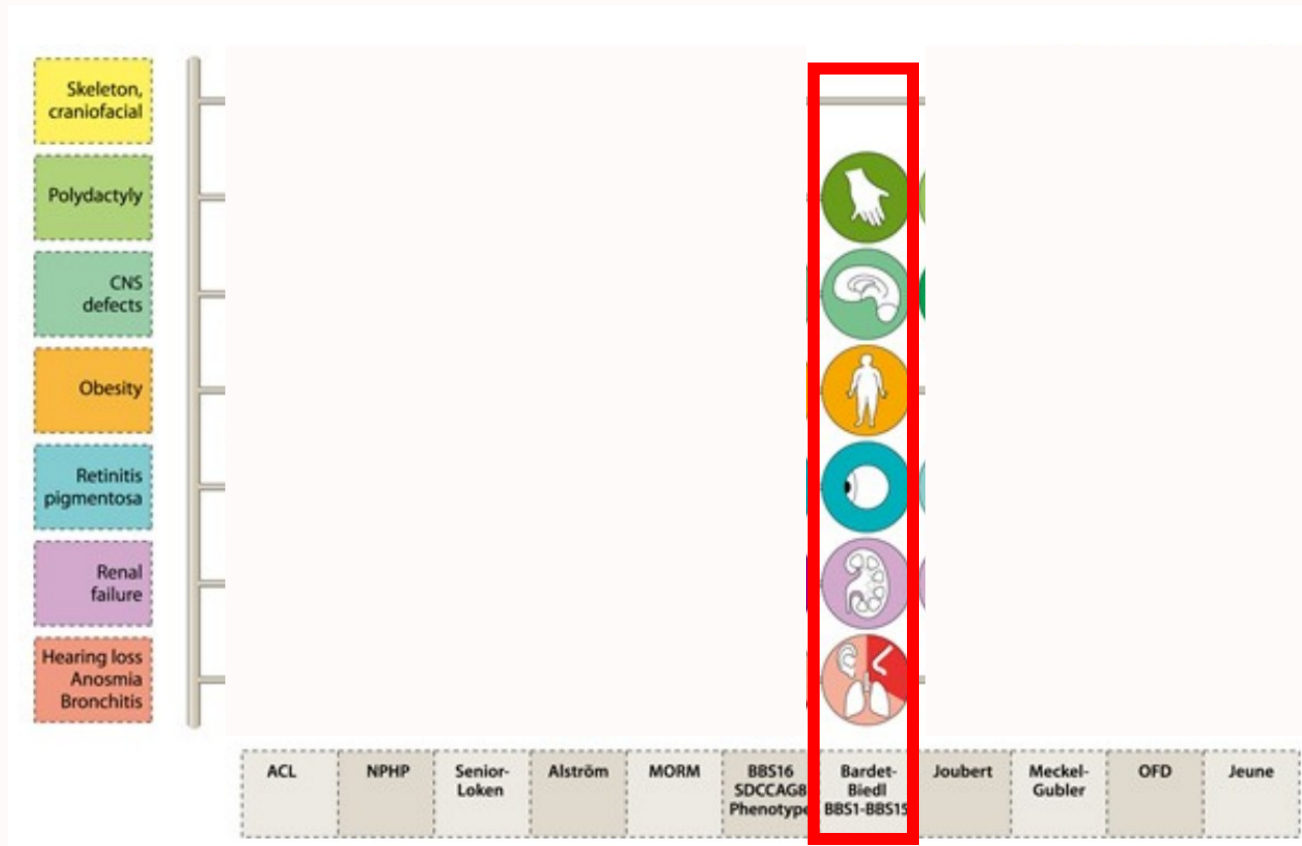
Autosomique récessif



# Syndrome de Bardet-Biedl.

Ciliopathie rare (1/150 000)

Autosomique récessif

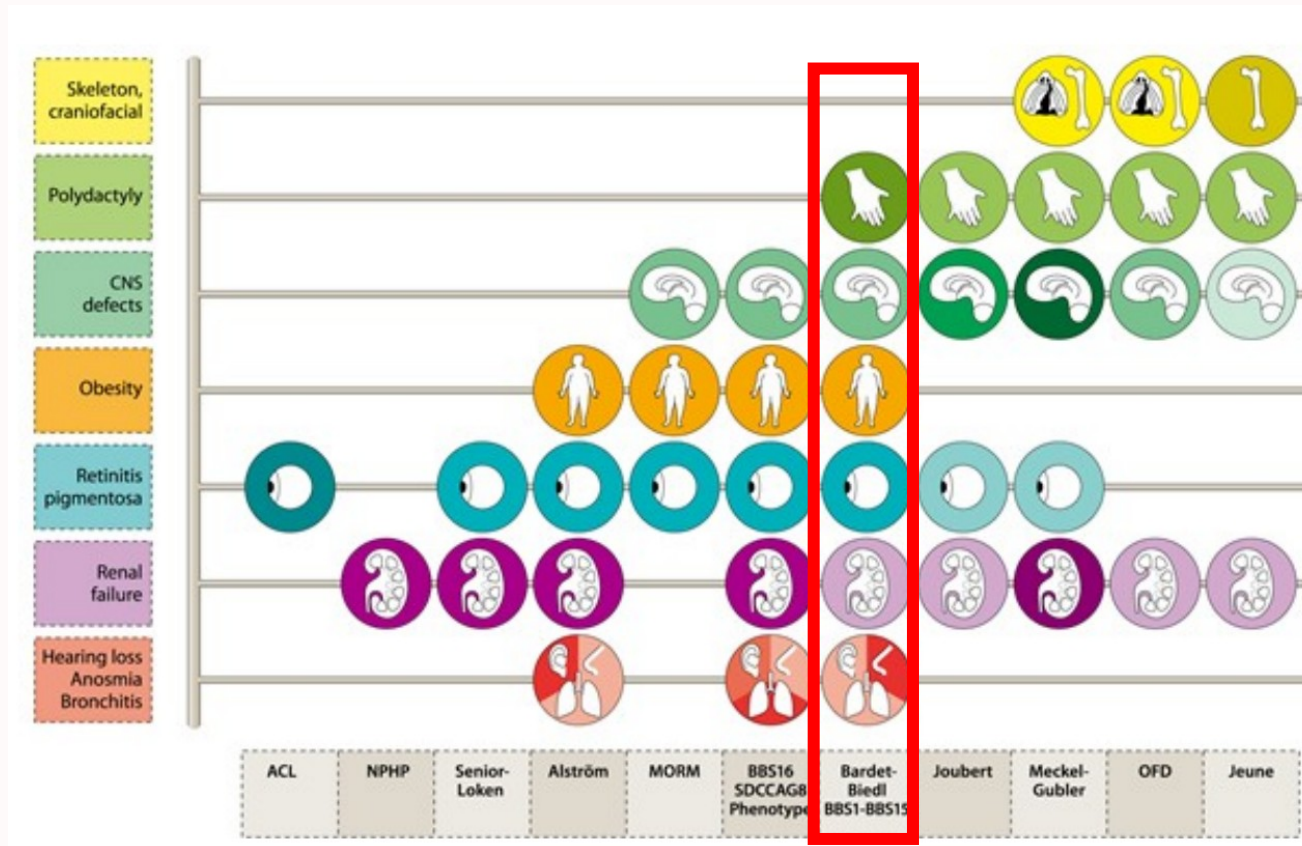


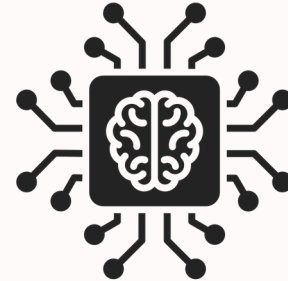


# Syndrome de Bardet-Biedl.

Ciliopathie rare (1/150 000)

Autosomique récessif



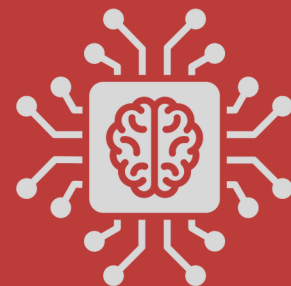


# Intelligence artificielle

Un système **autonome** capable de s'adapter après son déploiement **générant des résultats** influant sur son environnement **à partir des données qu'il reçoit.**

Adaptée de

EU Artificial  
Intelligence Act



# Intelligence artificielle



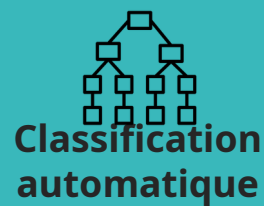
## Intelligence artificielle



Forêts  
aléatoires



Regression



Classification  
automatique

## Apprentissage automatique



## Intelligence artificielle

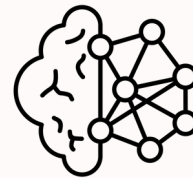


## Apprentissage automatique

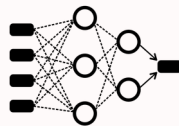


Transformer

s



## Deep Learning



FFN



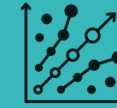
LSTM



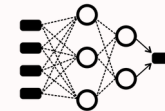
## Intelligence artificielle



## Apprentissage automatique



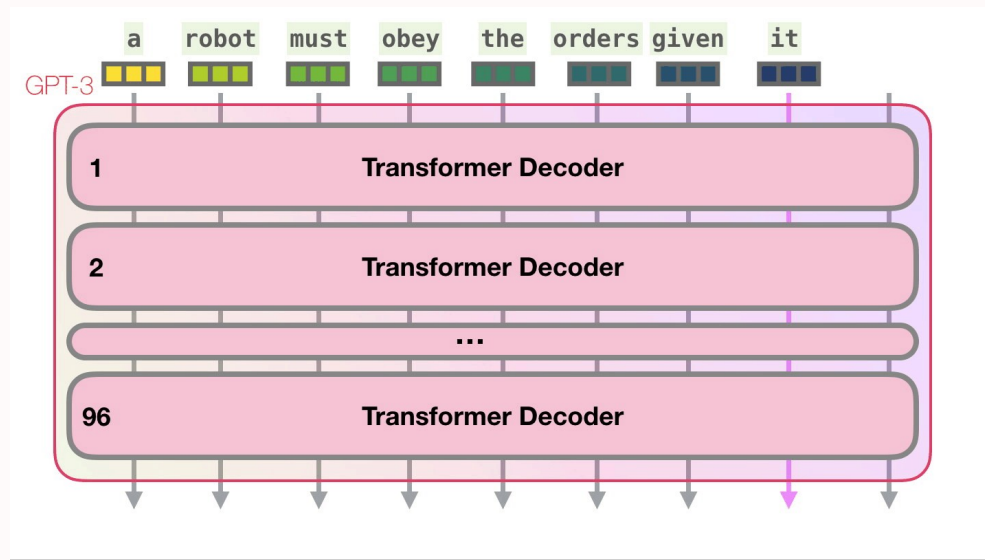
## Deep Learning





# Architecture Transformers

Apprentissage profond



<https://jalammar.github.io/how-gpt3-works-visualizations-animations/>

## (Self)Attention.

Ils sont attentifs à certains mots dans une séquence.

## Contexte.

Le contexte entier de la séquence est calculé.

## Prédiction.

L'architecture Transformers prédit la suite de la séquence.

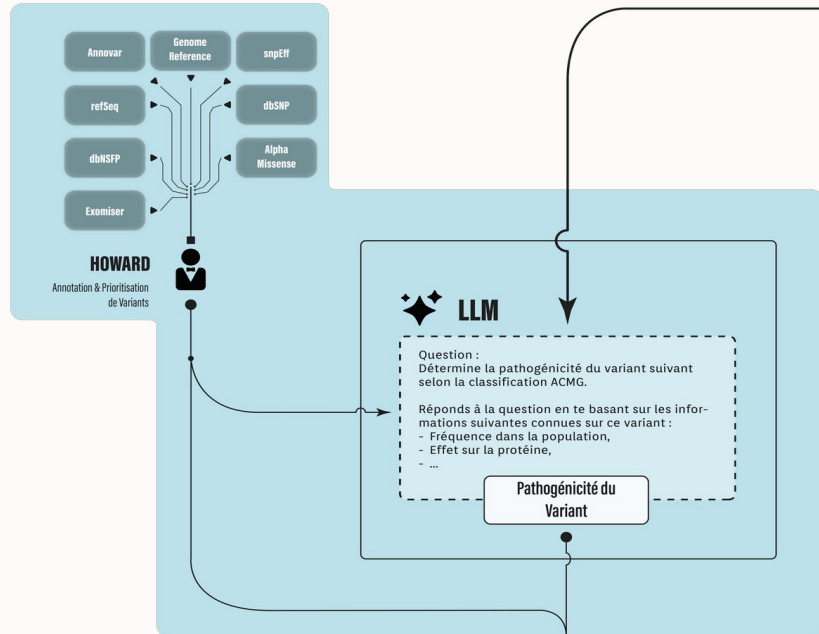
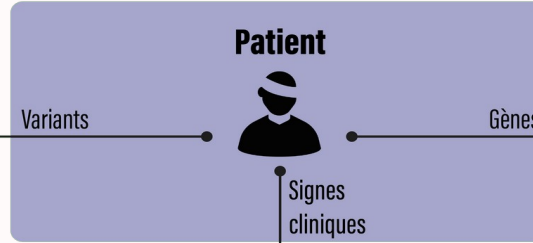
## Hallucination !

Le modèle cherche toujours à répondre donc il peut halluciner.

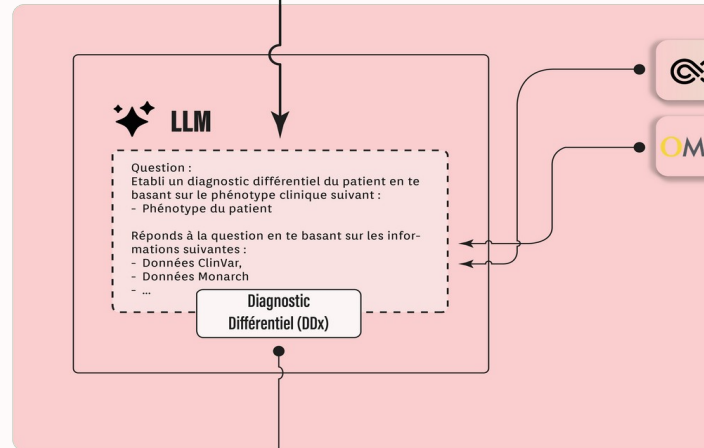


# Logigramme

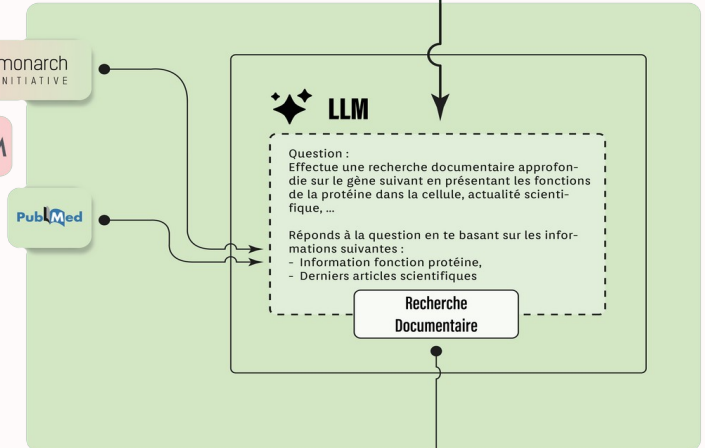
## Planner



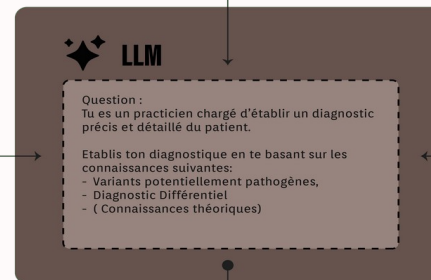
## Agent Génétique



## Agent Clinique



## Agent Biologique



## Agent Décisionnaire

Diagnostic Complet



## Modèles LLM : Mixtral 8x7B

- 46,7 milliards de paramètres.
- 8 experts de 7 milliards de paramètres
- Modèles quantisés : moins performants mais plus accessibles.



Modèle	Bits utilisés par paramètre	Taille	RAM Max requise
Mixtral 8x7B Q3	3	20.36 Go	22,86 Go
Mixtral 8x7B Q4	4	26,44 Go	28,94 Go
Mixtral 8x7B Q5	5	32,23 Go	34,73 Go
Mixtral 8x7B Q6	6	38,38 Go	40,88 Go
Mixtral 8x7B Q8	8	49,62 Go	52,12 Go

+ précis  
+ coûteux



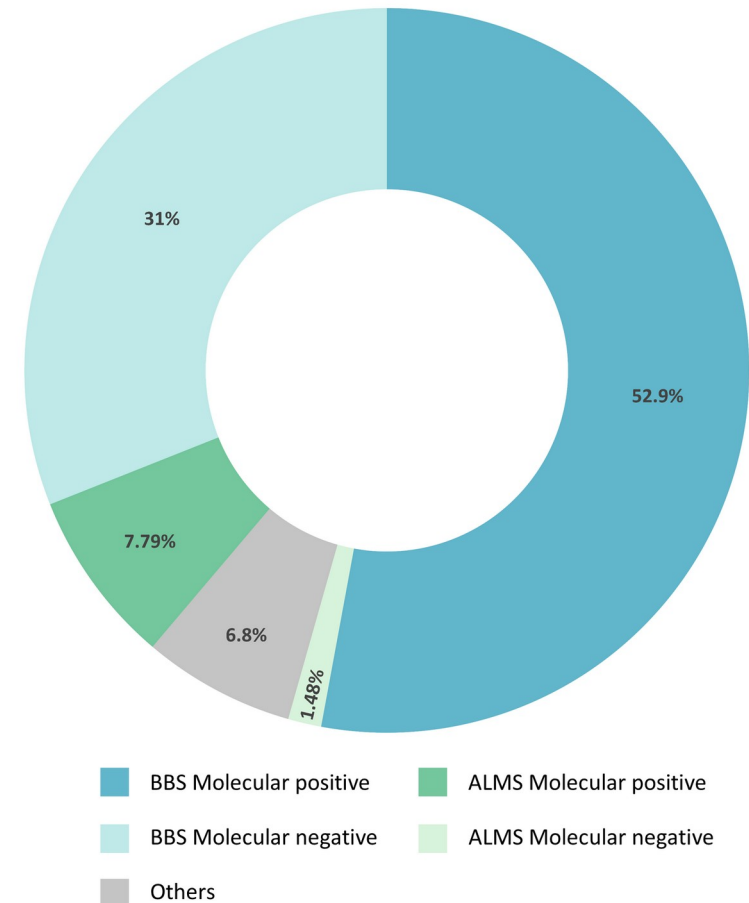


## Cohorte Syndrome de Bardet-Biedl (BBS)

- 846 patients adressés pour suspicion de syndrome Bardet-Biedl.
- Diagnostic clinique et / ou moléculaire connus.
- Test de l'agent clinique.

a.

Cohort subpopulations



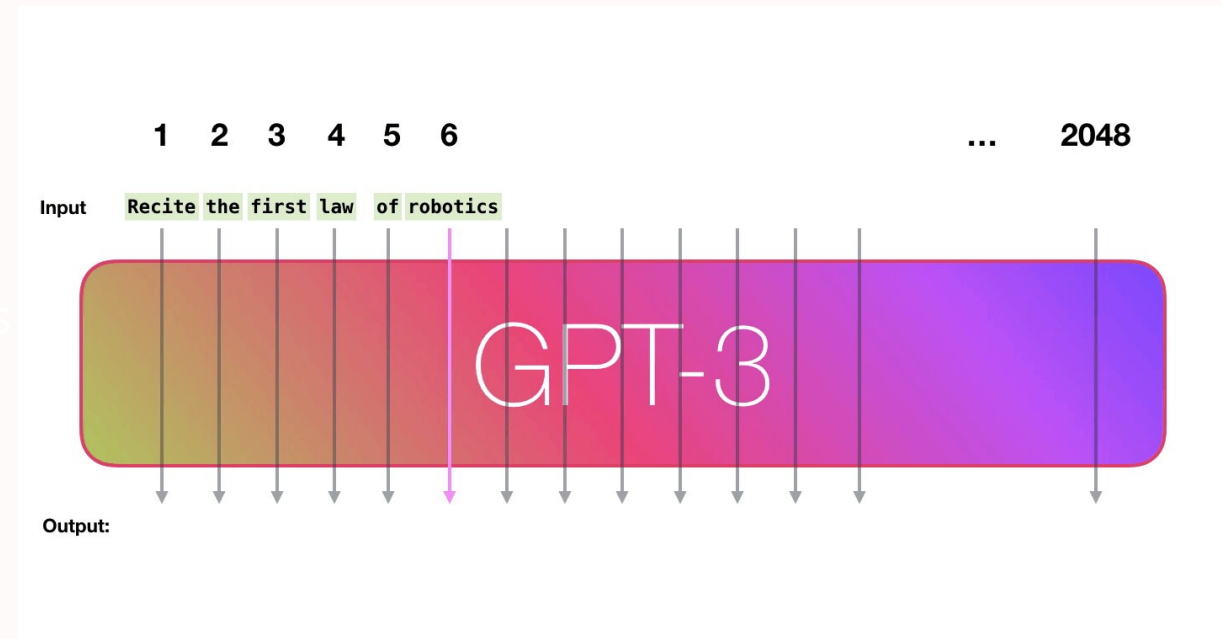


## La prompt

- Une instruction donnée au modèle pour atteindre l'objectif recherché : diagnostic différentiel, pathogénicité du variant, etc...
- Pour l'agent clinique :

Tu es **un clinicien** qui établit un **diagnostic différentiel (DDx)** [...] indique les 5 principales maladies génétiques rares possibles du patient [...] Tu répondras qu'en format JSON [...]

***Signes cliniques du patient***

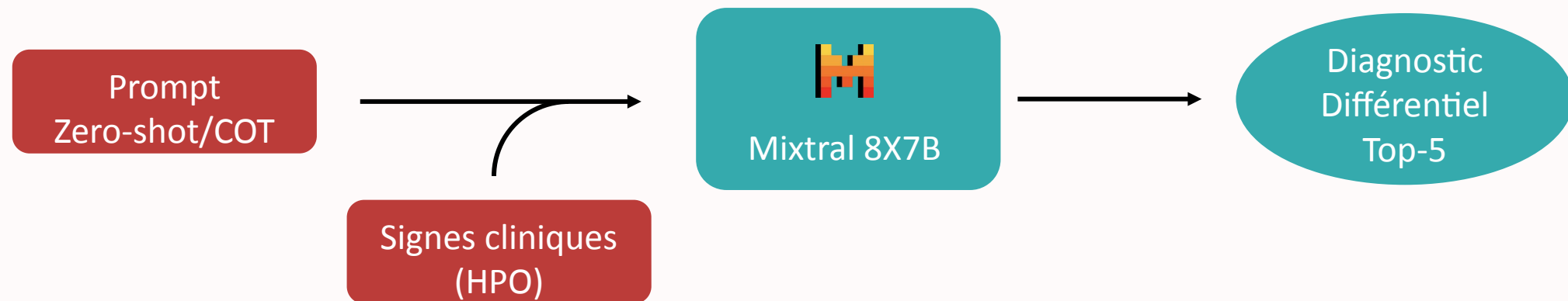


<https://jalammar.github.io/how-gpt3-works-visualizations-animations/>



## Résultats agent clinique

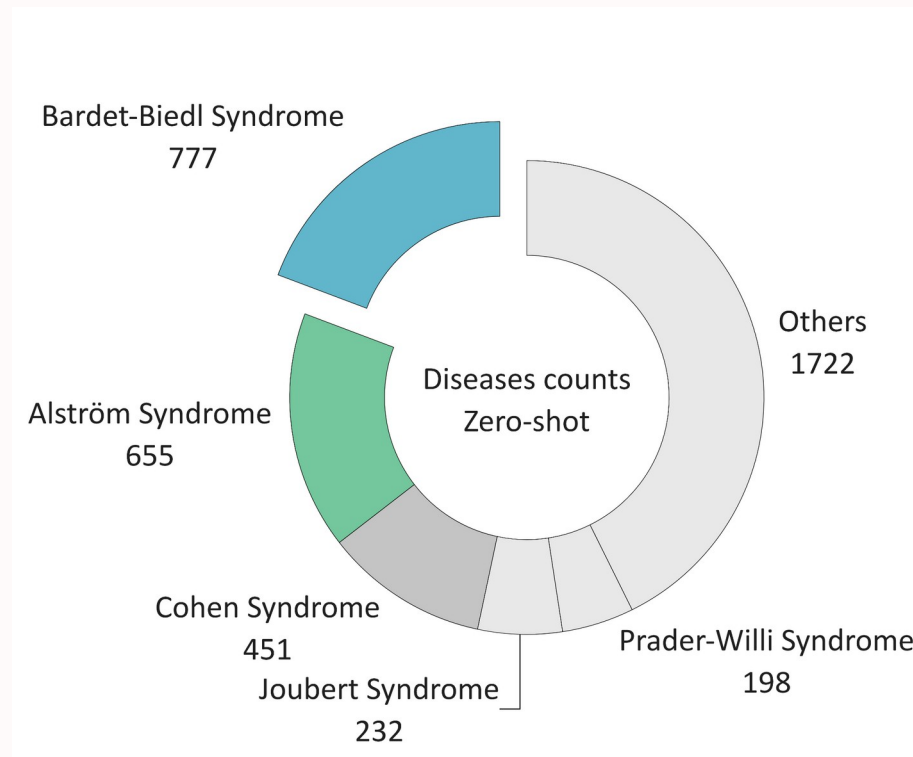
- Réalisation d'un diagnostic différentiel top-5.
- Evalué sur 2 prompts différentes : Zero-shot ou Chain-Of-Thought (COT).
- Matrice de confusion et analyse DDx.



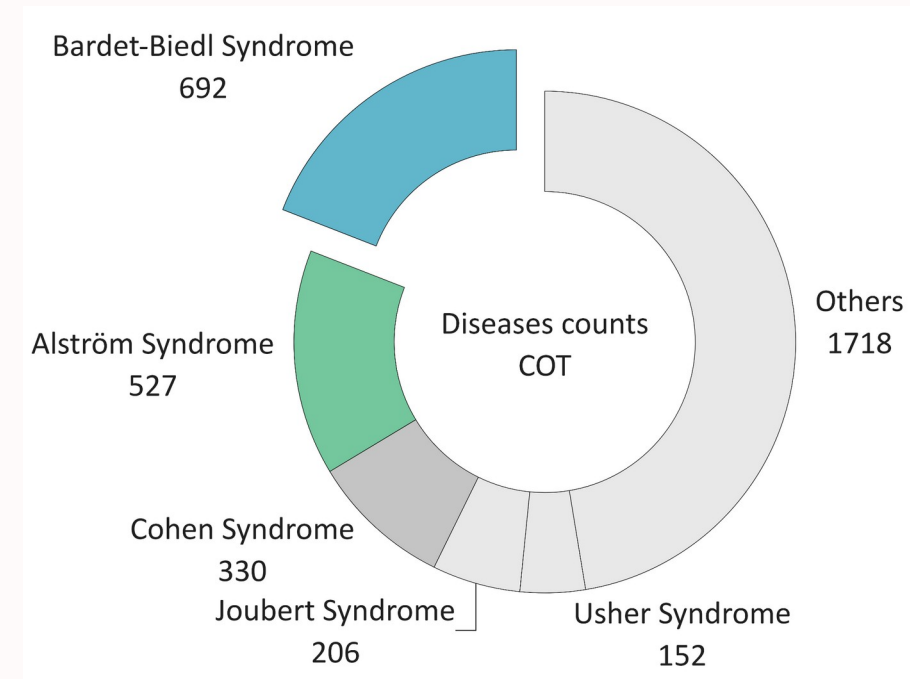
# Résultats agent clinique

Analyse de la cohorte complète.

**Zero-shot**



**COT**



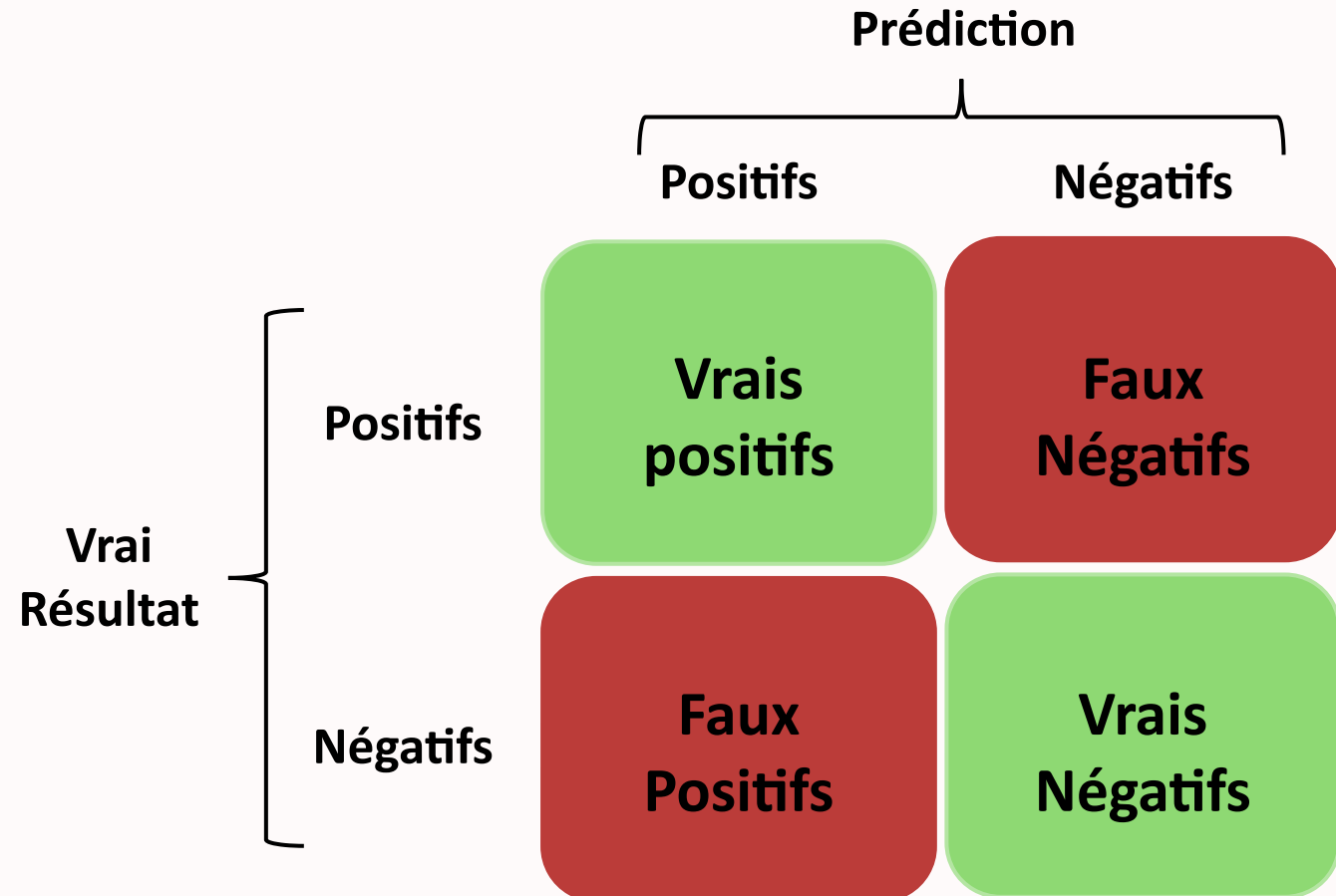
- Comptabilisation des top-5 pour 809 patients.
- Maladies citées présente des phénotypes similaires.



## Résultats agent clinique

Analyse des 809 DDx : Matrice de confusion.

- Vérification sur Bardet-Biedl est présent dans le top-5 = **échantillons positifs/négatifs prédits.**
- Vérification du diagnostic moléculaire = **échantillons positifs/négatifs réels.**





## Résultats agent clinique

Analyse des 809 DDx : Matrice de confusion.

- Vérification sur Bardet-Biedl est présent dans le top-5 = **échantillons positifs/négatifs prédits.**
- Vérification du diagnostic moléculaire = **échantillons positifs/négatifs réels.**

		Prédiction	
		Positifs	Négatifs
Vrai Résultat	Positifs	424	3
	Négatifs	351	29

Mistral8x7B\_Q6

# Résultats agent clinique

Analyse des 809 DDx : Matrice de confusion.

	Zero-Shot					CoT				
Modèle	Q3	Q4	Q5	Q6	Q8	Q3	Q4	Q5	Q6	Q8
Tps d'exécution / patient (s)	7	5	6	10	94	29	21	25	31	229

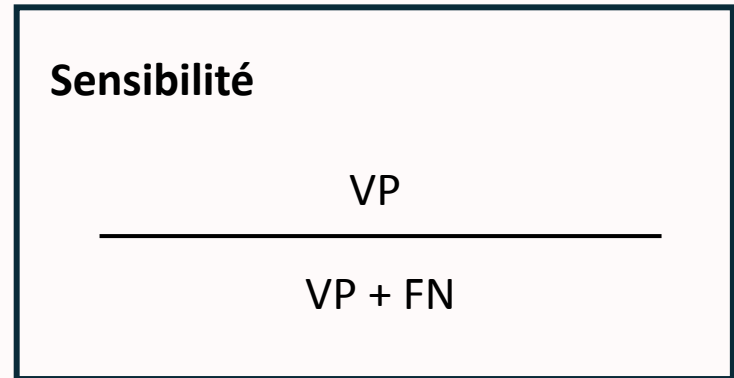
- Le zero-shot est plus rapide (2x)
- 5s / patient



# Résultats agent clinique

Analyse des 809 DDx : Matrice de confusion.

	Zero-Shot					CoT				
Modèle	Q3	Q4	Q5	Q6	Q8	Q3	Q4	Q5	Q6	Q8
Tps d'exécution / patient (s)	7	5	6	10	94	29	21	25	31	229
Sensibilité Top-1 (%)	81.3	86.2	<u>89.7</u>	89.5	89.9	69.9	81.5	82.2	84.7	84.0
Sensibilité Top-5 (%)	95.1	97.2	98.8	99.3	98.6	94.6	94.9	<u>99.0</u>	98.0	98.6



- Le zero-shot est plus rapide (2x)
- 5s / patient



# Résultats agent clinique

Analyse des 809 DDx : Matrice de confusion.

	Zero-Shot					CoT				
Modèle	Q3	Q4	Q5	Q6	Q8	Q3	Q4	Q5	Q6	Q8
Tps d'exécution / patient (s)	7	5	6	10	94	29	21	25	31	229
Sensibilité Top-1 (%)	81.3	86.2	<u>89.7</u>	89.5	89.9	69.9	81.5	82.2	84.7	84.0
Sensibilité Top-5 (%)	95.1	97.2	98.8	99.3	98.6	94.6	94.9	<u>99.0</u>	98.0	98.6
F1-Score Top-1 (%)	70.7	70.9	74.0	<u>74.2</u>	<b>74.5</b>	68.3	69.0	72.6	73.3	71.1
F1-Score Top-5 (%)	68.7	69.9	70.6	70.5	70.6	70.8	69.4	<b>72.4</b>	<u>72.0</u>	70.0

## F1-Score

$$2 \times \text{Précision} \times \text{Sensibilité}$$


---


$$\text{Précision} + \text{Sensibilité}$$

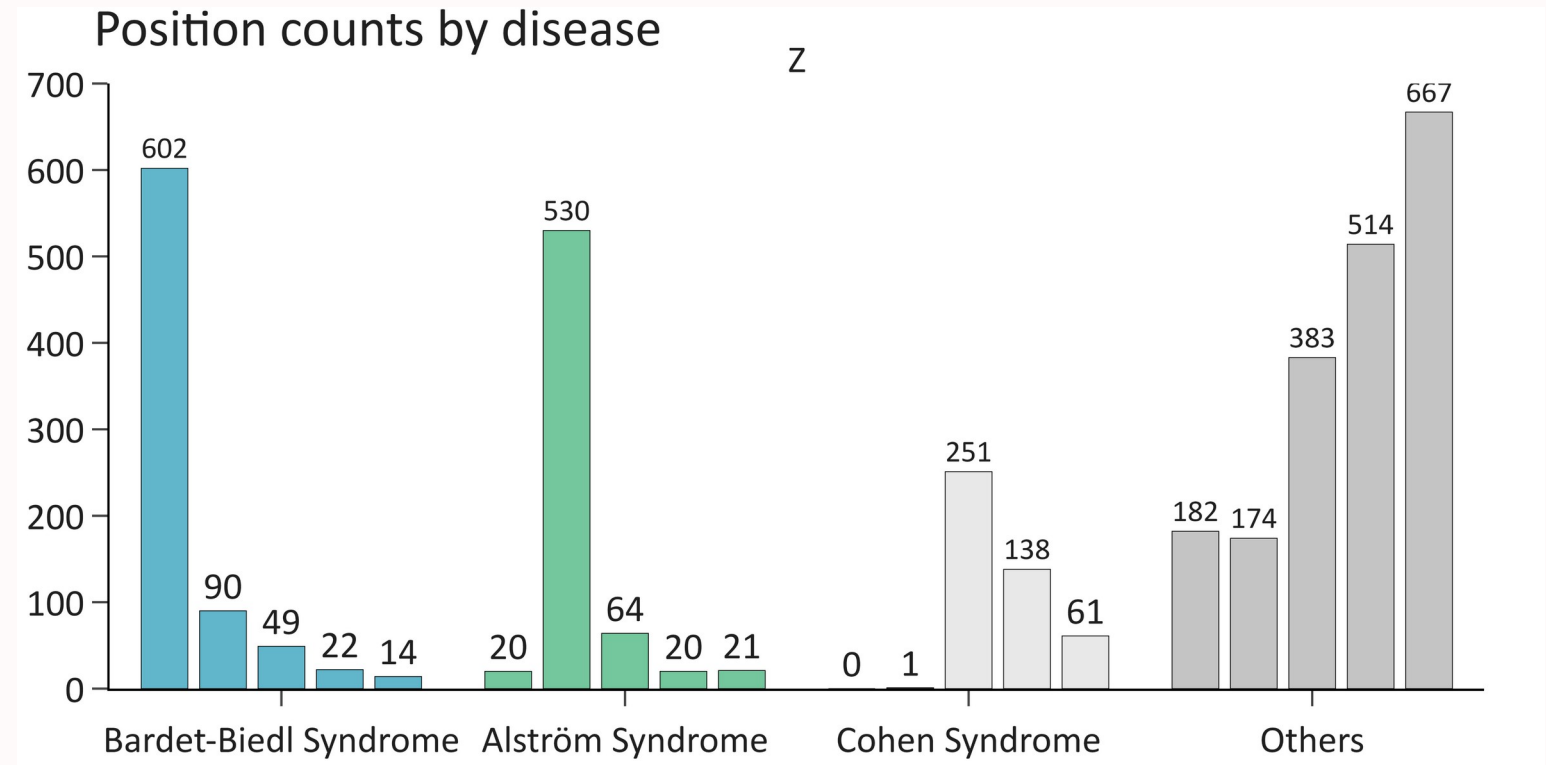
- Le zero-shot est plus rapide (2x)
- 5s / patient



## Résultats agent clinique

Décompte du nombre des positions 1,2,3,4 et 5 pour les maladies.

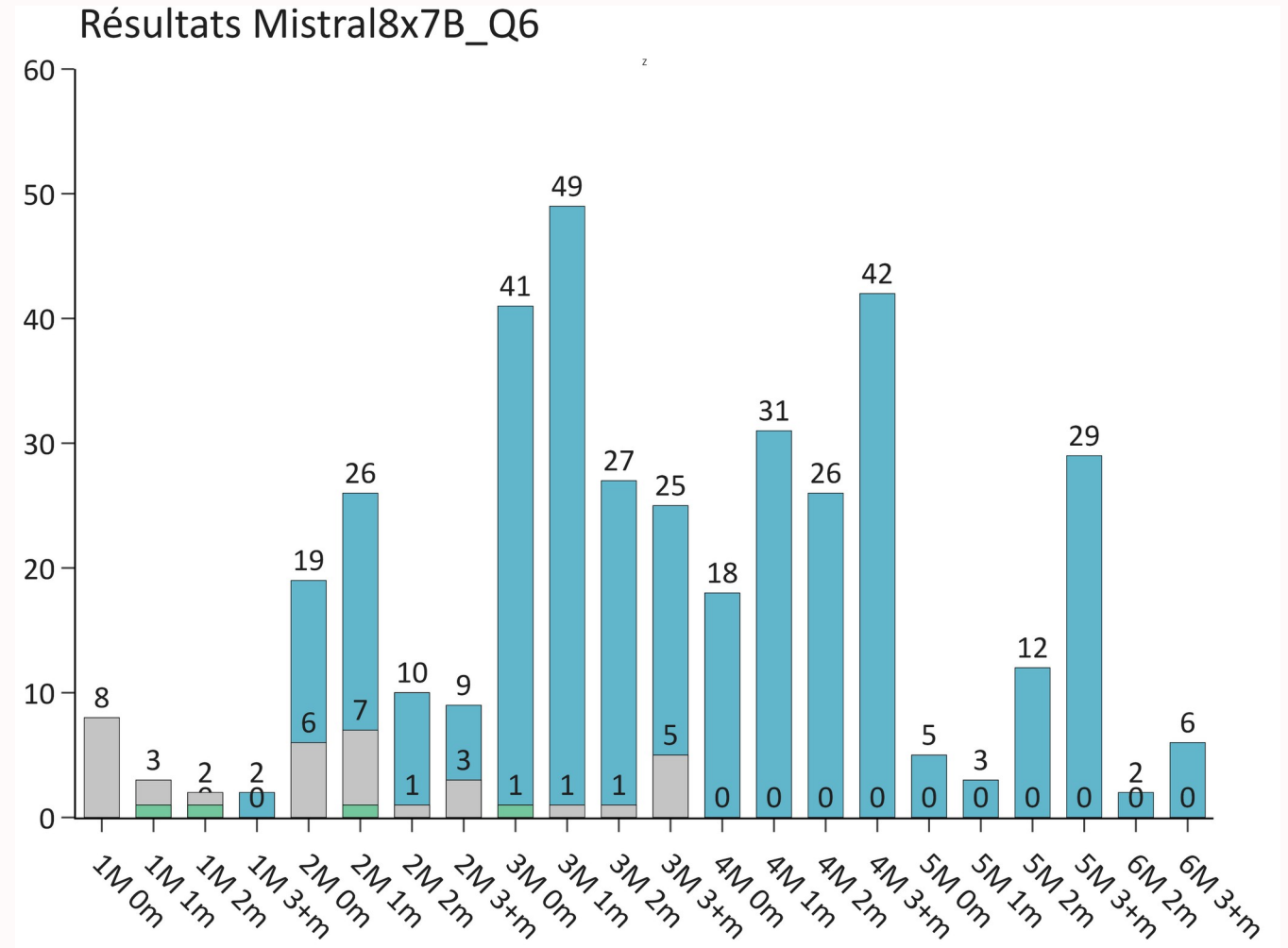
- Bardet-Biedl en 1<sup>ère</sup> position.
- Alström occupe la 2<sup>nd</sup> place.
- Réponses qualitatives de la part des modèles.



## Résultats agent clinique

Décompte du nombre des positions 1,2,3,4 et 5 pour les maladies.

- Critères cliniques aidant au diagnostic BBS.
- Plus un patient exprime ces critères plus BBS est en 1<sup>ère</sup> position dans le diagnostic.





## Conclusion & Perspectives

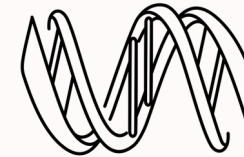


### AGENT CLINIQUE

**99%** de sensibilité Top-5  
**92%** de sensibilité Top-1

Bardet-Biedl principalement 1<sup>ère</sup>  
position.

**5s par patient.**



### AGENT GÉNÉTIQUE

Sur 6 variants : 5 variants  
correctement classés.

Sur 200 variants d'un même patient :  
**variant pathogène identifié.**

# Remerciement

Unité de Bio-informatique médicale appliquée au diagnostic

Laboratoire de génétique du NHC.

Dr. Caroline SCHLUTH-BOLARD

Dr. Jean MULLER et son équipe